# Explainable Deep Learning: A Field Guide for the Uninitiated

Year: 2020 | Citations: 414 | Authors: Ning Xie, Gabrielle Ras, M. Gerven, Derek Doran

---

## Abstract

Deep neural networks (DNNs) are an indispensable machine learning tool despite the difficulty of diagnosing what aspects of a model's input drive its decisions. In countless real-world domains, from legislation and law enforcement to healthcare, such diagnosis is essential to ensure that DNN decisions are driven by aspects appropriate in the context of its use. The development of methods and studies enabling the explanation of a DNN's decisions has thus blossomed into an active and broad area of research. The field's complexity is exacerbated by competing definitions of what it means "to explain" the actions of a DNN and to evaluate an approach's "ability to explain". This article offers a field guide to explore the space of explainable deep learning for those in the AI/ML field who are uninitiated. The field guide: i) Introduces three simple dimensions defining the space of foundational methods that contribute to explainable deep learning, ii) discusses the evaluations for model explanations, iii) places explainability in the context of other related deep learning research areas, and iv) discusses user-oriented explanation design and future directions. We hope the guide is seen as a starting point for those embarking on this research field.